Predict S&P 500 Movement at Market Opening Using Machine Learning Techniques

Eddie Chen '16; Advisors: Prof David Dobkin, Pat Chiacchiari (MFin '16)

PRINCETON UNIVERSITY

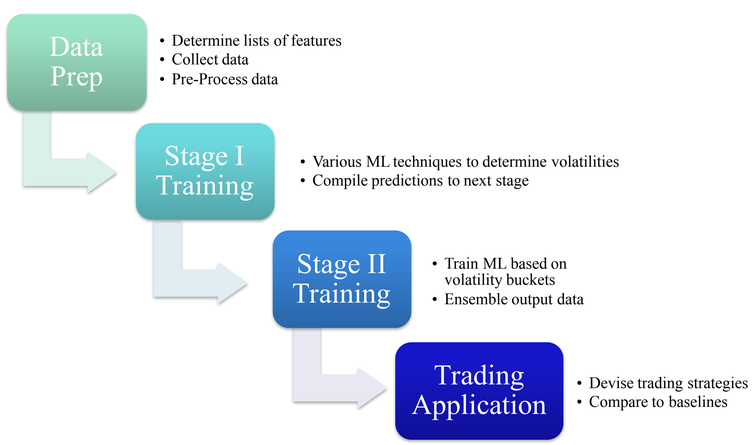
Motivation

- People love predicting stock markets for obvious reasons, and have tried and written extensively on various techniques from fundamental analysis, technical analysis, and sentiment analysis to moon phases and astrology. The difficulty lies in the enormous amount of data and the wide ranging variable types the stock markets churn out every day; while humans get easily overwhelmed and are prone to irrational biases, it seems a natural fit for machine learning.
- I was mostly motivated by my interest in this area of cutting edge industrial research, which is both intellectually stimulating and financially enticing.
- Moreover, few good public papers are available, making it a great learning opportunity.
- I was also inspired by Netflix Challenge which attracted a lot outstanding machine learning algorithm submissions.

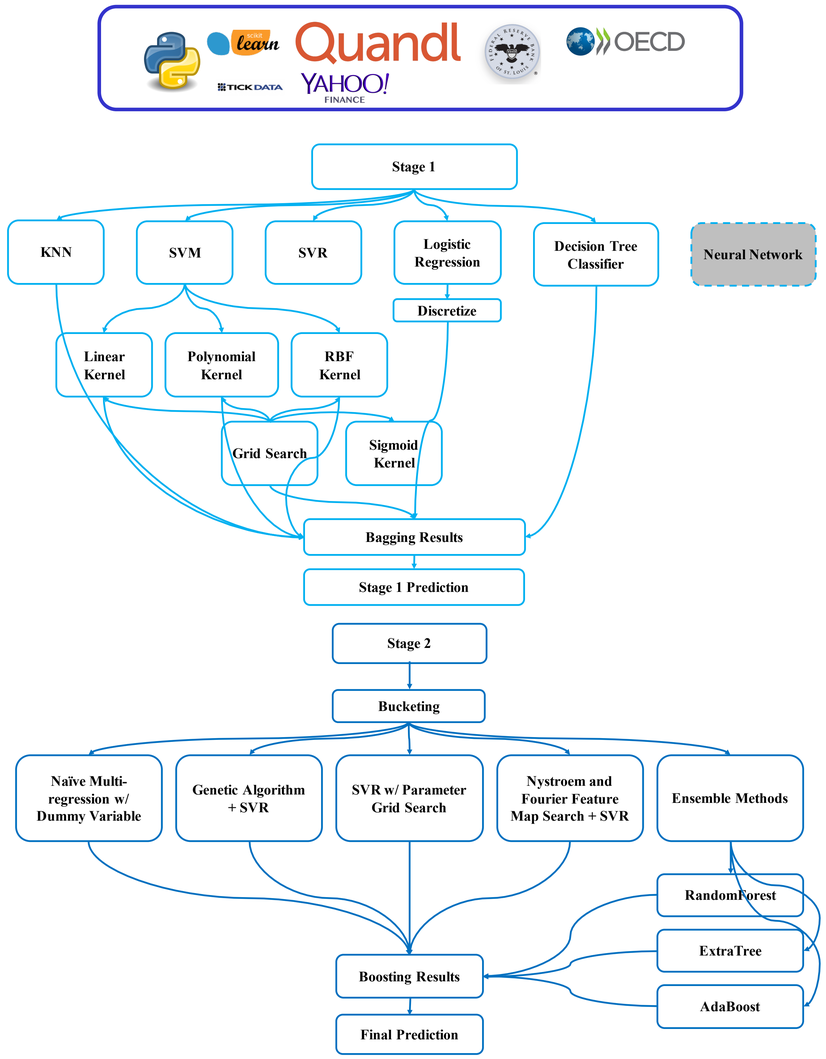
Financial Knowledge

- Researchers have long debated the validity of the **Efficient Market Hypothesis** (**EMH**) which states that share prices always incorporate all relevant information in real time, thus it is impossible to "beat the market" on a systematic basis, due to the random walk behavior of stock returns.
- Clearly, had even the weak form of the EMH been true, there would not be any need for projects like this one: because stock markets are efficient, it is pointless to predict any trends from public signals that are available now through computational means.
- I assumed in this project that the EMH doesn't hold, even in its weak form.
- Instead, I view the market as "sluggish", in that it doesn't incorporate all the information right away and when it does, it tends to over-react or under-react with the presence of white noise and systematic behavioral biases.
- But there are limits to arbitrage, therefore certain price distortions do not get arbitraged away.
- Moreover, according to the EMH, arbitrage is risk-less; but in reality it is risky because of fundamental risk, noise trader risk, margin calls on position and so on. As Keynes once famously quipped, "markets can remain irrational longer than you can remain solvent".

Proposed System



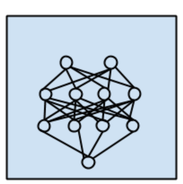
Two-Stage Approach

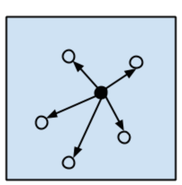


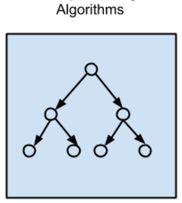
Limitations

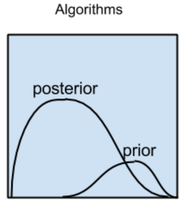
- **Small dataset – but S&P only came around in 1923!**
- **Only SVM-based techniques explored in detail**

Image Courtesy: Jason Brownlee, http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/


Deep Learning Algorithms


Instance-based Algorithms


Decision Tree Algorithms



Bayesian Algorithms

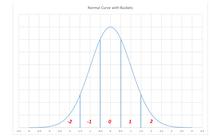
Project Future Work

- **Include Data from 1950 – Present**
- **Neural Network, Deep Learning**
- **Probabilistic Models**

Results

Stage I Results

Bucket	-2	-1	0	1	2
Expected Hit Ratio	6.70%	24.20%	38.20%	24.20%	6.70%



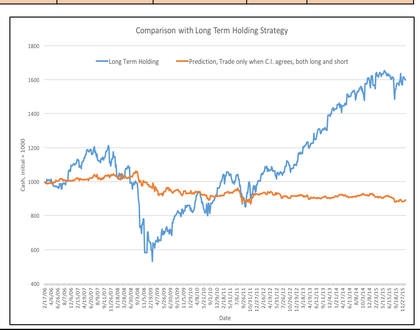
Algorithm	Score	Hit-ratio	RMSE	Note
K-NN (k=5)	0.533	53.3%	1.116	Standardized X_train
SVM w/ linear kernel, default params	0.570	57.0%	0.976	Standardized X_train
SVM w/ poly kernel, default	0.563	56.3%	0.974	Standardized X_train
SMV w/ rbf kernel, default	0.572	57.17%	0.925	Standardized X_train
K-means clustering (k=5)	-2991.945	8.33%~21.83%	2.218	Standardized X_train
SVM w/ GridSearch (rbf kernel, C=1, gamma = 0.0001)	0.567	56.67%	0.874	Standardized X_train
SVR, default setting	-0.036721728998	45.5% (discretized)	1.0886	Standardized X_train and y_train separately
Logistic Regression	0.572	57.17%	0.932	
SGDClassifier	0.567	56.67%	0.874	
CART Decision Tree	0.433	43.33%	1.231	
Bagging		57.8%		Bagging w/o SVR & SVR has a hit ratio of 60.8%
Averaging		57.5%		Averaging w/o SVR & SVR has a hit ratio of 61.5%

Stage II Results

Algorithm	Score	RMSE	Note
OLS Linear Regression	0.19	0.969	Standardized y
Random Forest Regression	-0.266	0.903	Standardized y
Extra Tree Regression	-0.335	0.922	Standardized y
Ada Boost Regression	-0.451	0.974	Standardized y
Grid Search SVR (rbf, C=1, gamma = 1)	0.007	0.799	Standardized y
Kernel Ridge SVR (alpha = 1.0, gamma = 1.0)	-0.018	0.808	Standardized y
GA, selected columns as ('Nikkei225Change', 5) ('GoldChange', 10) ('YenChange', 11) ('DAXChange', 17) ('FTSEChange', 18)	0	0.8	Linear Regressor as Fitness Function
GA, selected columns as ('YenChange', 11)	0.01	0.8	SVR as Fitness Function
Sign-aware averaging	-	0.8047	-

Trading Evaluations

Strategy	Initial Cash	Trading days	End Result	Return	Profitable Days
Trading only when C.I agrees, both long and short	$1,000	487	$890.20	-11.00%	143
Trading every day, both long and short	$1,000	600	$787.70	-21.20%	140
Benchmark, trading every day, long only, liquidate at end of the day	$1,000	600	$714.50	-28.60%	77



Acknowledgments

I would like to express my gratitude to Professor David Dobkin, for his patient guidance and great advice in picking and fine-tuning the project; to Pat Chiacchiari, for his awesome guidance on the finance side and many contributions such as advice on variable selections, market force awareness and trading strategy formulation. I would also like to thank IW 05 TA Ted Brundage for his help during office hours, and Yi Zhong (MIT '16) for her company and patience when I wanted to bounce ideas off someone. This project would not be possible without them.